COMP 5801H/COMP 4900 for Winter 2026
Generative AI and Large Language Models

Course Outline

Course Information

Instructor: Sriram Ganapathi Subramanian (`https://sriramsubramanian.com`)
Contact: sriramsubramanian@cunet.carleton.ca
Course Website: `https://brightspace.carleton.ca`
Lectures: 16.05 – 17.25 (Tue, Thurs) in-person (see Carleton Central or Brightspace)
Office Hours: 11:00 – 12:00 (Tue) in-person at HP 5360

You may book additional appointments (including Zoom meetings) via email. Online messages during this hour will be replied to promptly.

Discussion Medium: Students will be encouraged to use Piazza whenever possible (use this link to enrol: `https://piazza.com/carleton.ca/winter2026/comp5801h4900a`).
Required Tools: Python, Git, PyTorch or TensorFlow
Last Revised: Jan 8, 2025

Teaching Assistants

Contact info for your TA will be posted on Brightspace once the course starts.

Course Description

This course provides an exhaustive exploration of Generative Artificial Intelligence, with a focus on foundational mathematical principles, recent research breakthroughs, novel model architectures, and applications cutting across a wide range of domains. A variety of modern deep learning architectures, such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), Transformers, and Diffusion Models, will be covered in the course. The emphasis will be on practical aspects pertaining to the training and deployment of large models: data preparation, setting up training pipelines, evaluation of model performance, and strategies for improving training efficiency.

Prerequisites (for COMP 4900 only): COMP 3105 or COMP 3106; must have fourth year standing in BCS.

Textbooks (Optional)

Primary Textbook

- There is no required textbook for the course. Reading materials associated with each lecture will be provided.

Other Useful References

- I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.
- J. Eisenstein, *Introduction to Natural Language Processing*, MIT Press, 2019.

- Research papers from arXiv, NeurIPS, ICML, ACL, and EMNLP (reading list on Brightspace).

Course Format and Assessments

The course has two in-person classes every week. While the plan is for the lectures to be recorded, all students are highly encouraged to attend the lecture as each lecture will have ample scope for discussions and questions. The recordings are meant to be used if a medical or personal emergency prevents in-person attendance and for the purposes of revision of the lecture contents. If the in-person participation drops due to the availability of the recordings, the recording of lectures will be reconsidered.

The assessments in the course will be based on a combination of assignments and a final project. The assignments and the project will expect critical thinking, curiosity, and creativity from students. The final project will take the form of a conference paper, which may survey existing works or contain original research. Graduate students will complete the project solo, while undergraduates enrolled in the COMP 4900 section will work in groups of 2.

Inquiries

If you have a question (ex: clarification on readings, discussion about something said during class, questions about assignments), you should post them on Piazza so that your classmates can benefit from the discussion. If the question is about your assessments or situation, you may email the instructor or leave a private message on Piazza.

Please add COMP 5801 or COMP 4900 in your email subject line to ensure they are prioritized. Do not post code or assignment answers in the open or in course discussions. Questions about assessments will not be answered within 24 hours of the due date.

You may also schedule an appointment by emailing the instructor or assigned TA.

Topics Covered

Core Topics:

1. Recurrent Neural Networks (RNNs): Architecture and applications to sequential data
2. Attention Mechanisms and Self-Attention: Concepts and applications in transformers
3. Sequence-to-Sequence (S2S) Models: Encoder-decoder architectures for tasks like translation
4. Transformers: Introduction and training of transformers for NLP and other tasks
5. BERT and GPT: Understanding bidirectional and autoregressive LLMs
6. Reinforcement Learning from Human Feedback (RLHF): Alignment techniques in LLMs
7. Variational Autoencoders (VAEs): Dimensionality reduction and generative modeling
8. Generative Adversarial Networks (GANs): Adversarial training techniques
9. Diffusion Models (DDPM): Denoising diffusion probabilistic models for generative tasks
10. Ethics and Governance: Bias, fairness, and regulatory frameworks
11. Emerging Research: Multimodal models, autoformalization, and future directions

Advanced/Optional Topics (Subject to Available Time):

- Efficiency in Models: Knowledge distillation, efficient transformers, and sparsity methods
- Double Descent Phenomenon: Non-monotonic relationship between model capacity and generalization
- Inductive Biases in Deep Learning: Implicit biases in architecture and training
- Limitations of Classical Statistical Learning Theory: Why traditional ML theory falls short
- Multimodal Generative Models: CLIP, DALL·E, Whisper, and audio-text models
- Large-Scale Training and Distributed Systems: Parallelism, memory management, precision strategies
- Parameter-Efficient Fine-Tuning: LoRA, prompt tuning, and adapters
- Model Compression and Acceleration: Distillation, quantization, and pruning
- Advanced RLHF and Constitutional AI: Safety and reward modeling
- Advanced Inference Strategies: Sampling (Top-k, nucleus), decoding methods
- Retrieval-Augmented Generation (RAG): Knowledge-enhanced generation techniques
- Curriculum and Continual Learning: Learning efficiency and catastrophic forgetting
- Safety and Robustness: Adversarial defenses and alignment checks
- System Design and Deployment: Latency optimization and edge inference

Assessment Scheme

40%*	Final Project (due at the last day of classes)
60%	Assignments (3 assignments each worth 20%)

* Undergrads in COMP 4900 are to complete the project in groups of 2

Late Policy

The deadlines for assignments and projects will be strictly enforced, and no late submissions of assignments and projects will be accepted by default. If you need an exception for this, email the instructor at least 24 hours before the deadline. Note that more than one exception will not be made for the same student under any circumstances.

This policy accommodates unexpected circumstances such as technical and personal issues; therefore, no additional extensions will be granted (excepting accommodations provided by university policy).

Generative AI Tools

You are free to use any Generative AI Tool for your assignments or projects, but you bear responsibility for all your submissions. Since Generative AI tools may plagiarize from other sources on the web, you should be extremely careful while using them. They are also prone to producing incorrect information with exaggerated confidence. All prior work should be rigorously cited, and any plagiarism will be considered an academic offence.

Other academic boilerplate:

If you are unsure of the expectations regarding academic integrity (how to use and cite references, how much collaboration with lab- or classmates is appropriate), ASK your instructor. Sharing assignment or quiz specifications or posting them online (to sites like Chegg, CourseHero, OneClass, etc.) is considered academic misconduct. You are never permitted to post, share, or upload course materials without explicit permission from your instructor.

Academic integrity offences are reported to the office of the Dean of Science. Penalties for such offences can be found on the ODS webpage: `https://science.carleton.ca/academic-integrity/`.

uOttawa Graduate Students

- Brightspace access for University of Ottawa Students; please see information here: `https://gradstudents.carleton.ca/faculty-of-graduate-and-postdoctoral-affairs-access-to-brightspace/`
- University of Ottawa Students who need access to SCS IT resources such, as OpenStack and Nextcloud, must submit a request to SCS Tech Support SCS.Tech.Support@cunet.carleton.ca. The request must be sent from their @cmail.carleton.ca email address and the email should say which resource is required and for which course (including section).

Undergraduate Academic Advisors

The Undergraduate Advisors for the School of Computer Science are available in Room 5302HP; or by email at scs.ug.advisor@cunet.carleton.ca. The undergraduate advisors can assist with information about prerequisites and preclusions, course substitutions/equivalencies, understanding your academic audit and the remaining requirements for graduation. The undergraduate advisors will also refer students to appropriate resources such as the Science Student Success Centre, Learning Support Services and Writing Tutorial Services.

Graduate Academic Advisors

The Graduate Advisors for the School of Computer Science are available in Room 5302 HP; or by email at grad.scs@carleton.ca. The graduate advisors can assist with understanding your academic audit and the remaining courses required to meet graduation requirements.

SCS Computer Laboratory

Students taking a COMP course can access the SCS computer labs. The lab schedule and location can be found at: `https://carleton.ca/scs/tech-support/computer-laboratories/`. All SCS computer lab and technical support information can be found at: https://carleton.ca/scs/tech-support/. Technical support staff may be contacted in-person or virtually, see this page for details: https://carleton.ca/scs/tech-support/contact-it-support/.

University Policies

Academic Accommodations

Carleton is committed to providing academic accessibility for all individuals. Please review the academic accommodation available to students here: https://students.carleton.ca/course-outline/.

Academic Integrity

Student Academic Integrity Policy. Every student should be familiar with the Carleton University Student Academic Integrity policy. A student found in violation of academic integrity standards may be sanctioned with penalties which range from a reprimand to receiving a grade of F in the course, or even being suspended or expelled from the University. Examples of punishable offences include plagiarism and unauthorized collaboration. Any such reported offences will be reviewed by the office of the Dean of Science. More information on this policy may be found on the ODS Academic Integrity page: https://science.carleton.ca/students/academic-integrity/.

Plagiarism. As defined by Senate, "plagiarism is presenting, whether intentional or not, the ideas, expression of ideas or work of others as one's own". Such reported offences will be reviewed by the office of the Dean of Science. More information and standard sanction guidelines can be found here: https://science.carleton.ca/students/academic-integrity/. Please note that content generated by an A.I.-based tool could be plagiarized material (students should be careful while using these tools in their assignments and projects).

Unauthorized Collaboration. Senate policy states that "to ensure fairness and equity in assessment of term work, students shall not co-operate or collaborate in the completion of an academic assignment, in whole or in part, when the instructor has indicated that the assignment is to be completed on an individual basis".